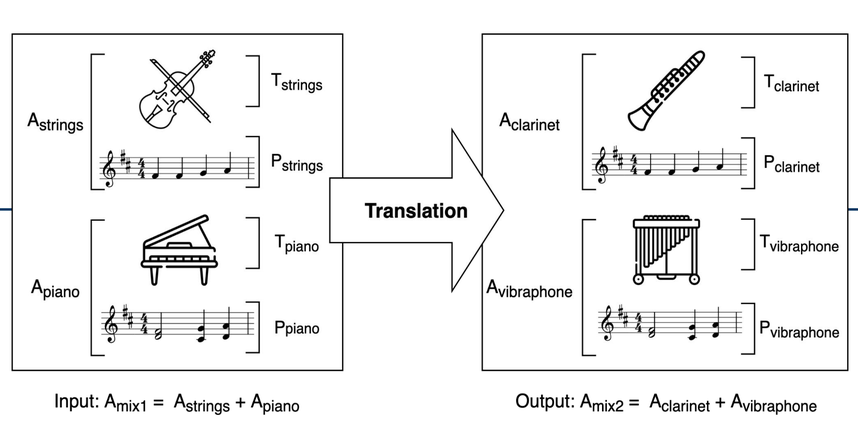
Music-STAR: A Style Translation System for Audio-based Re-instrumentation

Mahshid Alinoori, Vassilios Tzerpos
{mahshida, bil}@yorku.ca

Abstract

Music style translation aims to generate variations of existing pieces of music by altering the style-related characteristics of the original piece while content remains unchanged. Music style translation on raw audio has been investigated and applied to single-instrument pieces. In this work, we introduce **Music-STAR**, the first audio-based translation system that translates the existing instruments in a piece into a set of target instruments without using source separation.



An example of multi-instrument translation

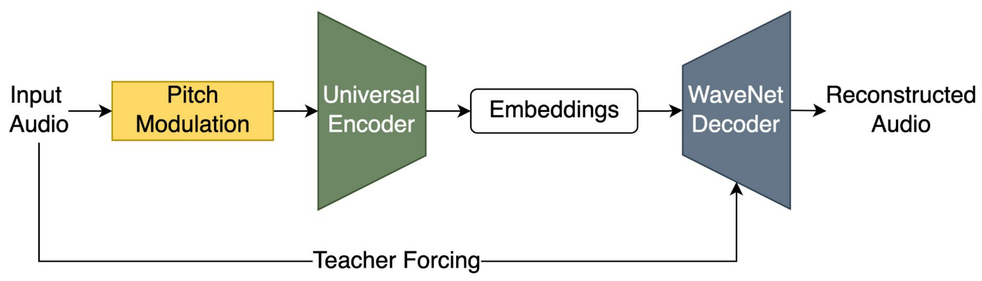
StarNet Dataset

We have created an audio dataset containing pieces that are composed of two instruments, and are performed with two domains: **Strings-Piano** and **Clarinet-Vibraphone**.
Every piece in the dataset is represented by both mixtures and their corresponding stems.

Methodology

We present several baseline solutions and then propose Music-STAR, which is built upon the **WaveNet autoencoder:**

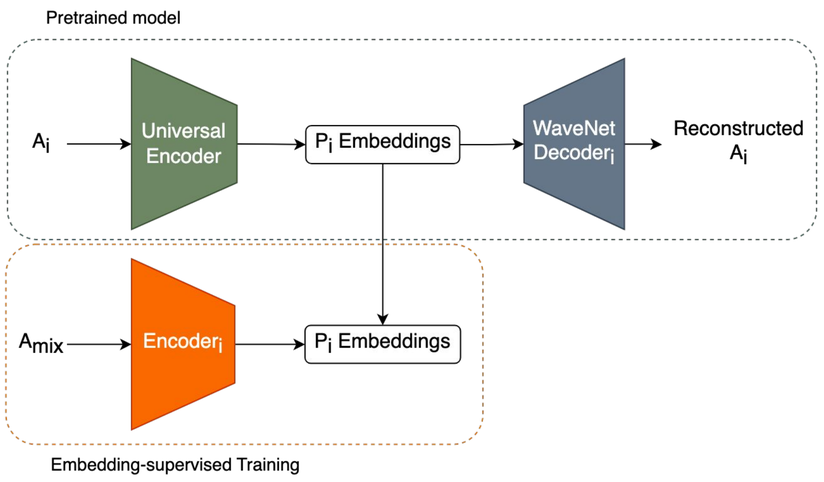
- **Single-instrument Translation Pipeline** applies an existing single-instrument translation model, the universal translation network, to the pre-existing stems and mixing the outputs.



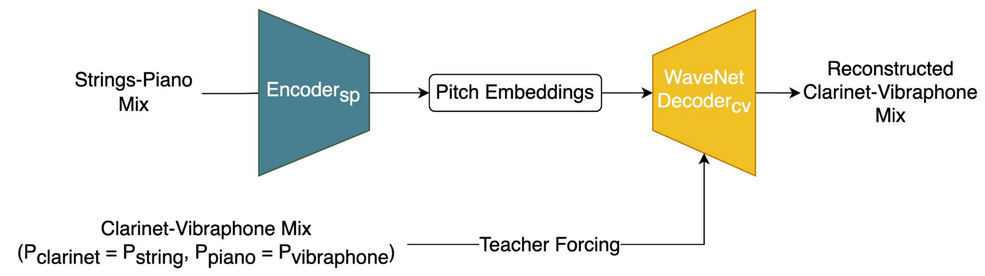
The universal translation network is trained as a denoising autoencoder using teacher forcing

- **Separation-based Translation Pipeline** applies
  a. **Demucs** source separation model to isolate the stems in the input mixture,
  b. The **single-instrument translation** network to the isolated tracks and mixes the outputs.

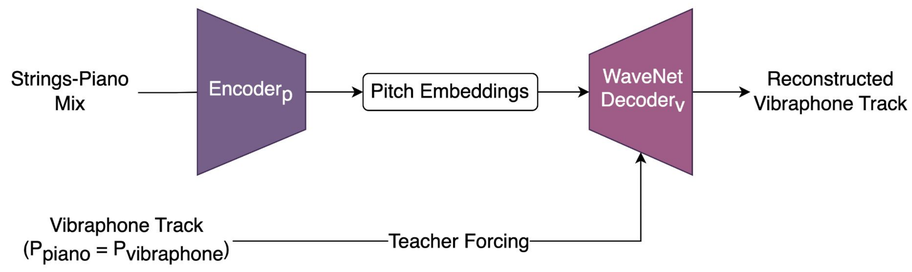
- **Embedding-supervised Method** is performed by training an encoder to generate the same embeddings as the universal encoder for every instrument present in an audio mix separately:



- **Mixture-supervised Music-STAR** is performed by training the WaveNet autoencoder using teacher forcing technique where the target audio mixture is the decoder's input during training:



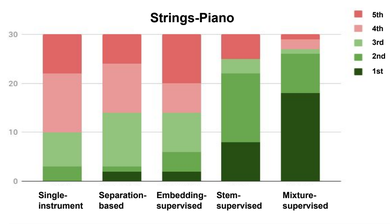
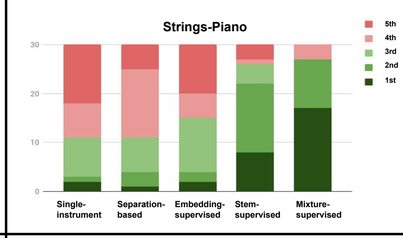
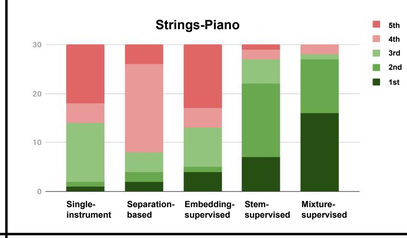
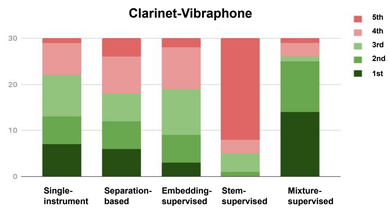
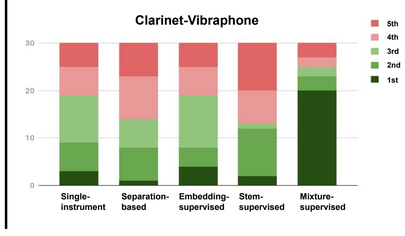
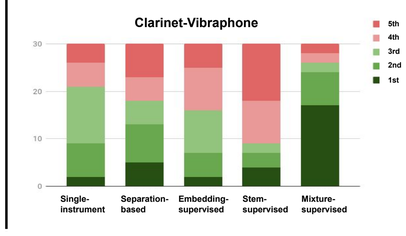
- **Stem-supervised Music-STAR** is performed by
  a. Training two autoencoders using teacher forcing technique where a single target instrument track is the decoder's input during training for each
  b. Mixing the outputs of each autoencoder during inference



Evaluation

Subjective and objective evaluations compare the re-instrumentation methods on three criteria:
- Content preservation
- Style fit
- Audio quality

Ranking of methods provided by subjective evaluation

Content Preservation	Style Fit	Audio Quality
		
		

Summary of method scores based on subjective rankings and objective metrics

Method	Subjective			Objective	
	Content	Style	Quality	Content (Jaccard)	Style (Cosine)
Single-instrument	166	150	153	0.371	0.483
Separation-based	165	147	159	0.392	0.474
Embedding-supervised	161	157	149	0.350	0.472
Stem-supervised	154	190	183	0.323	**0.699**
Mixture-supervised	**254**	**256**	**256**	**0.426**	**0.698**

Conclusion

Music-STAR tackles multi-instrument translation without applying explicit source separation to the input mixtures. We explored a variety of possible solutions based on the WaveNet autoencoder, and finally reached a successful mixture-supervised method which is outperforming the baselines.

Scan the QR code to listen to the audio samples.